Analyzing Sub Par Kickstarter Data

1. Given the provided data, what are three conclusions we can draw about Kickstarter campaigns?

With the provided data we can draw the following conclusions:

* 1. There are substantially more plays on Kickstarter than any other Sub-Category\*
  2. Projects that launch in December are less likely to succeed than projects launched in other months
  3. Given a sub-category, it is more likely to have all projects in that sub-category to be in 1 State (Successful, live, failed, or canceled) than have a mix of states.\*

\*These are not actually true, but appear to be with this data

1. What are some limitations of this dataset?

One limitation to this data set is that the location is not as narrowed down as it could be since Kickstarter projects in the US have city and state.

Other data that would be awesome is information on the different backer levels that each project has and how many backed at each level.

An important thing to analyze for each project would also be how much funding that the project had throughout the campaign. Some projects fund in 1 hour where others fund on the last day.

I think that my points (a) and (c) on question 1 address two other limitations. This data set appears to be curated as opposed to randomly sampled. As a veteran of the board game industry, I know that board game projects not only make up a significant amount of Kickstarter’s revenue, but also data with 80 projects showing that 100% of the projects fund is incredibly unlikely.

1. What are some other possible tables and/or graphs that we could create?
   1. I think that Average Donation by Category/Sub-Category would be something worth while to look at as well as backers\_count.
   2. Length of campaign vs. Number of backers and Average Donation
   3. Another comparison I’d like to make vs. Category and Sub-Category is spotlight
2. Bonus: Use your data to determine whether the mean or the median summarizes the data more meaningfully.

Because of the incredible right-skew on these data sets, median summarizes the “center” of the data much better.

1. Bonus: Use your data to determine if there is more variability with successful or unsuccessful campaigns. Does this make sense? Why or why not?

There is more variability in the count of backers in successful campaigns. I think that this makes sense because campaigns that didn’t have enough interest tend to have very little interest, which means the data ends up grouped very close to the 0 backers end.

**Additional Comments:** Steps 1 and 3 do not align with modern data visualization practices. Multiple colors in a column like that is frankly disgusting and would be much better represented quantitatively with a simple bar graph. Nothing stands out if everything stands out. Red Green Blue is also not colorblind friendly.